

データの分析で学んだ公式からの考察

— 回帰直線と相関係数の関係 —

統計学 数理研究班

2023 年 11 月

1 研究背景

数学 I で学んだデータの分析の単元で、相関係数 r は

$$-1 \leq r \leq 1$$

と解説してあるが、証明が与えられていない。分布図の傾向を最もよく表す 1 次直線として、回帰直線を導入してるが、理論的な解説がない。

2 研究目的

これらの理論的な根拠を学習するとともに、相関係数と回帰直線との関係性について研究を行った。

主に次の 3 点について研究を行った。

1. $-1 \leq r \leq 1$ の証明.
2. 回帰直線の方程式を求める.
3. 相関係数と回帰直線の傾きとの関係を研究.

3 $-1 \leq r \leq 1$ の証明

次の t の 2 次関数を考える ($f(t) \geq 0$).

$$\begin{aligned} f(t) &= \frac{1}{n} \sum_{k=1}^n \{t(x_k - \bar{x}) - (y_k - \bar{y})\}^2 \\ &= s_x^2 t^2 - 2s_{xy}t + s_y^2 \end{aligned}$$

$f(t) = s_x^2 t^2 - 2s_{xy}t + s_y^2$ は、 $f(t) \geq 0$ であるから、係数について

$$D/4 = s_{xy}^2 - s_x^2 s_y^2 \leq 0$$

したがって

$$\frac{s_{xy}^2}{s_x^2 s_y^2} - 1 \leq 0 \quad \text{ゆえに} \quad r^2 - 1 \leq 0$$

よって $-1 \leq r \leq 1$

4 回帰直線

2 つ変数 x, y からなる n 個データ

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

について、 $\frac{1}{n} \sum_{k=1}^n (y_k - ax_k - b)^2$ を最小にする定数 a, b を求める.

$$F(a, b) = \frac{1}{n} \sum_{k=1}^n (y_k - ax_k - b)^2$$

とおき、 F を a, b について偏微分すると

$$\begin{aligned} \frac{\partial F}{\partial a} &= -\frac{2}{n} \sum_{k=1}^n (y_k - ax_k - b)x_k \\ &= -2\{E(xy) - aE(x^2) - b\bar{x}\} \end{aligned}$$

$$\begin{aligned} \frac{\partial F}{\partial b} &= -\frac{2}{n} \sum_{k=1}^n (y_k - ax_k - b) \\ &= -2(\bar{y} - a\bar{x} - b) \end{aligned}$$

F が最小となるとき、 $\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$ であるから

$$aE(x^2) + b\bar{x} = E(xy), \quad a\bar{x} + b = \bar{y}$$

したがって

$$a = \frac{E(xy) - \bar{x}\bar{y}}{E(x^2) - \bar{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2}, \quad b = -a\bar{x} + \bar{y}$$

このとき、直線 $y = ax + b$ 、すなわち

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

を回帰直線という.

5 結論

回帰直線によって、ある値 x_i に対する y の予測値を計算できる。また、相関係数 r と回帰直線の傾き a の符号が一致する。